Healthcare: Persistent vs Non-Persistent

Abstract: We analyse a dataset which features various factors which affect the persistency of a drug. This is a Binary Classification problem, where we build models that can best predict patient persistency. This data-driven approach will aid pharmaceuitcal industries to optimise their strategies to improve patient outcomes. Various factors include “anti-depresseants and mood stabilisers”, “vitamin d deficiency”, “chronic liver disease”, “low calcium intake” and more. After analysing the correlation between different factors, we built machine learning algorithms to classify patient data. We found that SVM models outperformed other models in terms of accuracy and precision, however Logistic Regression had a higher recall. Due to the nature of the business question, we choose SVM as our final model since we prioritize Precision over Recall. Pharmaceutical companies can use this paper to build their own predictive models for persistent analysis.

# Introduction

Pharmaceutical companies have the responsibility to prescribe drugs to those with medical conditions and deficiencies. Prescriptions last for a certain amount of time, and patients usually must stick with their user-selected drug therapy plan. Persistency Analysis focuses on the continuation/ discontinuation of the drug for different patients and their prescriptions. Results can be used to improve medical adherence, which is a priority for health systems worldwide.

Big pharmaceutical companies in the world, like GSK (GlaxoSmithKline) and Sanofi can benefit from producing more drugs which are persistent and improving non-persistent drugs to increase sales. They can also create modifications of the drug for different audiences ((e.g. 18-35 may react differently to 35+), with the latter being proven to having consistent persistent results) over drug X. We can introduce a method of altering prescriptions so that we do not over or under prescribe.

The ABC Pharmaceutical company is aiming to understand the persistency of a drug per the physician subscription. To achieve this, they have approached an analytics company to automate a process of identification, which we will be doing in this project.

A group of vials on a conveyor belt

Description automatically generated

Figure : Drug Prescription

We begin by harnessing data pre-processing, data cleaning and data transformation techniques on the data before analysing the trends between different factors and their correlation. After, we build models leveraging machine learning algorithms to predict the likelihood of a patient complying to the prescribed treatment regimen. This can be further enhanced by hyperparameter tuning and optimisation, which we will do before finding the results. A comparative analysis will be done on the correlation threshold (we choose factors based on the strength of their correlation) and the different models. Lastly, we choose the final model for this binary classification problem which will have the highest accuracy.

# Analytical Questions

## What are the main factors in driving persistency of this drug among patients?

We use a correlation matrix to see which factors have the highest relevancy to the persistency of the drug. We look to analyze and generate insights from previous patient data for the drug and create visualizations of these.

## Which machine learning models are best suitable for predicting persistency?

Construction of different machine learning models with optimisation and comparison is they focus of this task. Ultimately, we want to build a model which can help pharmaceutical companies predict whether or not the drug will be effective and the patient will complete his/her prescription based on their medical history. We look at SVM, KNN, Naïve Bayes, Decision Trees, Random Forest and Logistic Regression for predictions.

Comparative analysis of Machine Learning algorithms will be done with accuracy, precision, recall, f1-scores and ROC-AUC curves to choose the final model. We will produce the confusion matrix of the final model.. Also a comparison of all results to a baseline model (Dummy Classifier) will be completed.

## Relationships between different factors for Persistency

### Age x Persistency

### Gender x Persistency

### Race x Persistency

### Region x Persistency

We also aim to produce many different data visualizations presenting the relationships between different factors for persistency. We look to fit different regression models and plot their results. We also analyze any outliers in numerical variables presenting the data points in a box plot.

## Hypothesis Statement

### SVM will be the optimal algorithm for this dataset

The SVM model is particularly useful with datasets with many features and fewer values because it can differentiate data points better with its hyperplane in a higher dimensional space.

# Data (Materials)

This is a binary classification problem. The data contains several Boolean variables, with either “Y” – True, and “N” – False for different medical conditions. We will analyze which one of these conditions affects drug persistency the most.

**Data Storage Location:**

<https://www.kaggle.com/datasets/harbhajansingh21/persistent-vs-nonpersistent>

A white table with black text

Description automatically generated

Figure : Tabular Data Intake Report

There are 50 Boolean variables, 16 Categorical variables and 3 Numerical variables. We convert the Boolean and Categorical variables into Numerical variables, so that we can compare correlations to the target variable “Persistency” for data preparation. The target variable is categorical, and we proceed by using a one-hot encoder, to change these values to either 0 or 1. We highlight the issue with this encoder is the that it leads to a larger feature space

There is notable class imbalance, between some variables in the dataset. We have 2135 non-persistent recordings compared to 1289 persistent recordings. Also, there are 3230 females compared to 194 males and 3148 patients from the Caucasian race, and 276 patients from other races. We will mitigate class imbalance by oversampling persistent recordings to get a balanced dataset. Furthermore, we transform all Boolean variables into 0 or 1 (0: False, 1: True) for our final models. We encode our numerical categorical models into 0, 1, 2 ,3 … respectively for the number of categories for machine learning.

Due to the large number of factors, this will increase the sophistication of the analysis, insights, and models to create a more reliable predictor model. We can also explore plenty of relationships between factors to give us a better understanding. However, there are imbalances between some Boolean variables, with only 11 patients having oestrogen deficiencies and 14 patients being immobilized, so we have to consider the bias of the data for these factors. More data will increase the reliability of the models.

# Analysis

## Analysis of General Factors (Age, Region, Number of Risks, Age Brackets, Ethnicity)

We begin our analysis with a bar chart visualizing the Gender statistics across the Target Variable. We can see the massive imbalance in Gender within the study, with only 194 male subjects compared to 3230 females. If we consider ratios, there are 94.03% of females who are persistent and 94.51 % of females who are non-persistent. This is a miniscule difference, so no conclusions can be drawn here.

A graph with green squares

Description automatically generated

Figure : Bar Chart showing Persistency across Gender

After converting all the data to numerical variables, we begin by analysing the correlation between the target variable “Persistency” and the other variables. Figure 4 shows that the highest number of people who have their drug persistency recorded are in the Midwest, which also has the highest number of non-persistent cases. The Southern region finishes second, with number of subjects but has a higher proportion of persistent cases. The West region has the highest proportion of persistency, with 44.49%, followed by the Northeast (42.25%) then the South (39.65%) and lastly the Midwest region (32.45%).

A graph of different colored squares

Description automatically generated

Figure : Bar Chart showing Persistency across Region

A close-up of a pie chart

Description automatically generated

Figure : Pie Chart showing differences in Proportions across Regions for Persistent and Non-Persistent Drug Cases

Figure 5 emphasises the insights from Figure 4, showing the largest proportion and number of persistent and non-persistent cases are in the Southern region. We can see that there are only 59 patients in the “Other/Unknown” region with 25: 34 persistency ratio.

A graph of the number of perseverance

Description automatically generated

Figure : Analyzing Persistency across the Number of Risks

We move on to analyse a new independent variable, “Count of Risks”. The “Count of Risks” variable highlights the number of possible “risks”/ contributing factors which correspond to the patient sticking with their prescription or not. In this study, most patients have only 1 risk. There are only 2 subjects with 7 risks, and 6 with 6 risks. Both non-persistent and persistent variables follow a similar distribution, with no real difference other than a higher proportion of non-persistent cases had 0 risks.

A graph of different colored bars

Description automatically generated

Figure : Persistency across different Age Bucket

Majority of people within this study are over 75, with a high proportion of non-persistent cases being reported in this bucket. Again, like in Figure 6, Figure 7 shows a similar trend across different categories. There are only a small number of cases where the age of the user is below 55 (166) showing some class imbalance. The older you are, the more likely you are to be prescribed with this drug. We can see from figure 7, that age only has a slight impact on persistency. The older you are, the more likely you won’t persistent although there is not enough evidence to back this claim.

A close-up of a pie chart

Description automatically generated

Figure : Tobacco

Figure : Distributions across Specialties for Drug Cases

## Analysis of other Factors (Viral Vaccines, Vitamin D Deficiency, Tobacco Smoking)

A graph of different colored bars

Description automatically generated with medium confidence

Figure : Viral Vaccines, Vitamin D Deficiency

A higher number of non-persistent patients have not had viral vaccines, and if a patient has had a viral vaccine, they are more likely to be persistent. Perhaps, this is due to patients being more sensible and organized, so they are more likely to adhere. A higher proportion of patients which have Vitamin D Deficiency are persistent because they are required to take tablets daily to prevent bone health issues, muscle weaknesses and pain, cognitive impairment etc. A patient not being deficient doubles their chances of not being persistent based on this graph alone (there are of course, other contributing factors). Both factors in Figure 9 graphs have a contribution to the target variable.

A graph of smoking and tobacco

Description automatically generated

2780 patients do not smoke tobacco, compared to 644 patients which do. Out of Persistent cases, a higher proportion of these smoke tobacco, indicating a necessity for patients who have tobacco in their bloodstream to require more compulsive medication.

## Numerical Variable Analysis

A graph with a red line and blue lines

Description automatically generated

Figure : Diagnostic Plots of Numerical Variable "Count of Risks"

A graph with a line and a red line

Description automatically generated

Figure : Diagnostic Plot of Numerical Variable "Dexa\_Freq\_During\_Rx"

There are a significant number of outliers in the DEXA Scan plot, which is used to analyse bone density. This suggests that there are a lot of patients in this study with varying bone densities. We can see the distribution of the Count of Risks Variable in Figure 11 similar to Figure 6, with only two outliers in this feature.

## Models

We begin with model preparation of the data by creating a correlation matrix to highlight they key factors which affect the persistency target variable. We experiment with two correlation thresholds, 0.01 and 0.1, only choosing variables above these values.

### 0.1

We iterate through the columns within the Data frame and select the columns with a correlation greater than the absolute value of 1 (either positive or negative). There is 1 target variable and 31 independent variables in these models. Variables include Vitamin D Deficiency, Blood Glucose Records and Doralsiga.

A screenshot of a computer

Description automatically generated

Figure :Sample variables and Correlation for 0.1

### 0.01

We repeat the same, but with a lower threshold, hence there are more independent variables (43). New variables not seen in the 0.1 threshold data include Gender, Region and Count of Risks.

A screenshot of a computer

Description automatically generated

Figure : Sample Variables and Correlation for 0.01

A table with numbers and text

Description automatically generated

Figure : Model Accuracy Results

A table with numbers and text

Description automatically generated

Figure : Model Accuracies Precision, Recalls and F1-Scores

We can see that the Support Vector Machine models and the Logistic Regression model with a 0.01 threshold perform the best. We take these 3 models and optimise them using Grid Search. As you can see, Support Vector Machine models have higher accuracies, however Logistic Regression models have higher recall and f1\_scores. Higher accuracies in SVMs show they are better at overall classification, but it does not necessarily suggest how well they are performing on each class separately. Logistic Regression models have higher recall scores, so they have a better ability of capturing positive instances even if they are false positives. They are also capable of minimising false negatives. Higher precisions in SVMs also suggest they are effective in classifying a larger proportion of positive and negative instances correctly, and when the SVM predicts a positive class, it is more likely to be correct and not a false positive class.

A graph of a curve

Description automatically generated with medium confidence

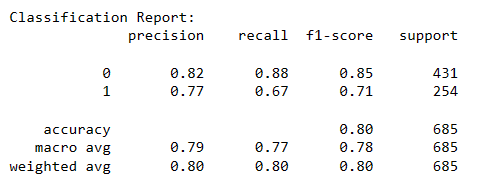
Figure : ROC Curves

ROC curves with SVC, SVC, Logistic Regressions and Dummy Classifier comparisons. We can see that the ROC (Receiver Operating Characteristic) giving a graphical representation used in Binary Classification The SVM models have a slightly higher true positive rate to false positive rate.

## Final Model

**A blue squares with white text

Description automatically generated**

****

# Findings, Reflections and Further Work

Our findings have shown that there are many factors which partially contribute to Persistency, and they Hypothesis statement mentioned above has been confirmed. SVM is the best model. We can see from the Final Model graphs above, that have been optimized using Grid Search to have the highest accuracy and precision. We prioritize models which have a higher precision than recall because predicting positive classes accurately is more important than detecting false negatives, as the result of a false negative in the real world will only be a small misclassification of a person not adhering to persistency (not a life-threatening) situation. Our final model has 381 true positives, 169 true negatives and 135 misclassifications with 50 false positives and 85 false negatives.

There are lots of outliers within the Dexa Frequency numerical data. As this is only one of the three variables, however it has a reasonable correlation with Persistency, we can consider removing these outliers.

Key issues with the data arise with class imbalance, and lots of features which have similar correlations coefficients. Models will be very complex, since there are lots of features which all contribute to Persistency, and to put these models into practice, it is difficult to find the right threshold to compromise between efficiency, overfitting, and accuracy. The more accurate models require a smaller threshold but take more time to train and are more expensive to implement.

To improve this further, we could consider an ensemble technique like stacking models to improve their overall performance creating a meta-model which can potentially outperform other models. This will introduce diversity and help capture different patterns within the data. Stacking can also improve generalization, reducing overfitting and balancing individual model biases. However, data leakage can occur due to improper cross-validation and the complexity of the model increases, requiring more computational resources and careful hyper-parameter tuning.

Further lines of enquiry include gathering more data to improve reliability of models, or stacking multiple models together (e.g. KNN-SVM). Further analysis can be done on other factors such as Fragility Fracture Recency and Injectable Experience. We could experiment with other thresholds such as 0.05 or just include all the variables in the model.

Further work can be done on getting more data from male patients, patients from Northeast and West regions, and monitor data on patients lower than 55. The exact age of the patient can be specified to give more evidence to potential trend in age and persistency. We can create regression models on features with the highest correlation to the target variable, to see how they impact individually.

Pharmacies can use this data to predict whether a customer will be persistent or not to the drug. We can alter prescriptions to drug to minimize company losses based on different factors and cater for different medical health conditions. More data can be provided for us to have a continuous persistency metric (e.g. 79.3% persistency) as opposed to just a binary classification. This can prevent the company from overprescribing drugs.

# Word Counts

|  |  |
| --- | --- |
| **Abstract** | 138/150 |
| **Introduction** | 286/300 |
| **Analysis Questions** | 270/300 |
| **Data** | 293/300 |
| **Analysis** | 953/1000 |
| **Findings, Reflections and Further Work** | 497/600 |